Running Head: Self As Object

Self As Object: Emerging Trends in Self Research

Jie Sui[1] and Xiaosi Gu[2]

1. Department of Psychology, University of Bath

2. School of Behavioral and Brain Sciences, University of Texas at Dallas

Correspondence:

j.sui@bath.ac.uk (J. Sui)

xiaosi.gu@utdallas.edu (X. Gu)

***Both authors contributed equally to this work.**

Abstract

Self representation is fundamental to mental functions. While the self has mostly been studied in traditional psycho-philosophical terms ('self as subject'), recent laboratory work suggests that the self can be measured quantitatively by assessing biases towards self-associated stimuli ('self as object'). Here we will summarize new quantitative paradigms for assessing the self, drawn from psychology, neuroeconomics, embodied cognition, and social neuroscience. We then propose a neural model of the self as an emerging property of interactions between a core 'self network' (e.g. mPFC), a cognitive control network (e.g. dlPFC), and a salience network (e.g. insula). This framework not only represents a step forward in self research, but also has important clinical significance, resonating recent efforts in computational psychiatry.

Key words: Self, other, objective measures, computational psychiatry, self network

Self as object

What is the nature of the self? This question has been central to a wide range of disciplines for centuries. In the early days of scientific psychology, William James proposed (1890) the existence of different aspects of the self – the 'Me' and the 'I' [1]. The former can be considered as 'self as object', whereas the latter refers to 'being an agent'. The 'Me' self is further composed of a physical, a social and a spiritual self [1] and the 'I' refers to being an agent. Sigmund Freud (1947) conceived of the self in terms of the *ego*, mediating between basic drives (the *id*) and social context/conscience (the *superego*) [2]. Although these early theories argued for the psychological reality of the self, researchers have more recently proposed that, similar to the concept of center of gravity [3], the concept of the self provides an explanatory narrative without serving as a mechanism that generates the narrative [4]. Following these arguments, much of the psycho-philosophical work has relied on subjective judgments and conscious evaluations of the self, what has been termed the 'self as subject' [5]. In this line of research, classic studies have used self-report measures to evaluate what people think and how they feel about themselves [6, 7]. Though useful, such methods are open to distortion through the demand characteristics of the social context and inferences made by participants about how to best present themselves. More recently, however, researchers have attempted to evaluate self processing by studying how associating stimuli to ourselves alters information processing (e.g. [8]). These studies focus on what we term the 'self as object' in mind and brain [9] and reveal several novel properties of self processing, including its ability to integrate perceptual, cognitive, and affective processes [5]. Here, it is worth distinguishing two definitions: self-related processing and self-referential processing [5,

10]. Christoff and colleagues argue that the self-experience of being an agent arises from self-specifying processing, and that both experiencing oneself as the 'I' and the 'Me' reflect self-related processing [10]. In contrast, Northoff argues that these aspects of self reflect self-referential processing rather than self-related processing, and that the latter refers to processing of any stimulus in relation to the brain [5].

We will distinguish between what we will term the 'self as object' [5] and 'self as subject' [1]; an important distinction between the two in this paper is that the former can be empirically operationalized through **empirical manipulation** and link to specific functional or neural processes (including both the 'Me' and 'I'), and thus could be used as a probe for psychopathology and as a model of how we make inferences about ourselves. By synthesizing these new findings from psychology, neuroscience, and **neuroeconomics**, and the emerging field of **computational psychiatry**[11] (see Glossary), the aim of this paper is to put forward a new framework for quantifying and objectifying the self in the context of mental function and dysfunction. We then discuss the neural circuits involved in self processing and propose a neural model of the self as an emerging property of the interactions between three brain networks: a core 'self network' centered in the medial prefrontal cortex (mPFC), a cognitive control network consisting of the dorsolateral PFC (dlPFC) and posterior superior temporal sulcus (pSTS), and a salience/affective network involving the insula and amygdala. This neural framework has broad applications in neuropsychological and psychiatric disorders characterized by alternations in self processing.

New paradigms for measuring the self

In this section, we review new paradigms related to self processing from experimental psychology, neuroeconomics, embodied cognition, and social neuroscience, and argue that this work can provide a new quantitative approach to understanding the self in mental function and dysfunction.

Self in experimental psychology: perceptual matching.

The past ten years or so has seen a burgeoning experimental psychology literature in which researchers have investigated how referring a stimulus to the self alters information processing, such as perception, attention, memory, visual awareness and decision making [12-16]. Notably, there is evidence that self-reference acts as a form of 'integrative glue' which can either enhance or disrupt performance, depending on the task context, a process often referred to as **integrative self** [17]. For example, in face perception tasks where subjects are asked to classify faces as self, friend, or stranger, classification of self faces is faster than that of the faces of other people due to enhanced feature processing and the integration of facial features into configurations [18]. Furthermore, self-reference can also bind different stages of information processing [19-21]. For example, one event-related potential (ERP) study examined the effects of facial cues on the orienting of visual attention. There was both an enhanced attention-related component N1 and a reduced decision making component P3 for self relative to others' faces [21]. This suggests that heightened attention to the self is coupled with reduced

uncertainty in decision making, which may then enhance binding between different stages of processing.

These effects on face processing are supported by findings from *perceptual matching tasks* (**Fig. 1A**) using neutral geometric shapes [16, 20]. In a typical experiment, participants associate a personal label (e.g., stranger, friend, you) to a neutral shape (e.g. triangle, circle, square) [16] and they immediately favor the shape associated with themselves compared to other shapes (**Fig. 1B**). This effect suggests that perceptual judgments are enhanced by tagging external stimuli with internal self concept (i.e., referred by the personal label). Furthermore, this behavioral effect is subserved by enhanced connectivity between the mPFC (associated with self representations) and the pSTS (reflecting attention to external stimuli) [19, 20] (**Fig. 1C** and **1D**) [20]. Northoff (2016) has argued that one of the key functions of the mPFC is to reflect representations of ourselves [22]. The pSTS is associated with sensory integration and self/other processing [22, 23]. Neuropsychological data additionally show that the breakdown of the vmPFC and the insula leads to the loss of self-biased responses in perceptual matching [24].

========== Fig. 1 approximately here ==========

This paradigm has been subsequently used to assess whether self-reference affects perceptual integration between stimuli. After forming associations between one personal label and two shapes, participants were asked to identify single or pairs of shapes as referring to the self or to a friend in a categorization task [8]. When the shapes referred to the self, there was a large benefit from presenting two, compared to one exemplar; that is,

there was a highly significant redundancy gain. Mathematical modeling showed that this redundancy gain was greater than what could be expected if there was independent processing of each self-shape exemplar [8]. Such benefit was not apparent for stimuli associated to a friend or to a monetary reward [25]. These results suggest that self-reference enhances integration between self-related stimuli. Perceptual matching tasks have also been used to compare the role of reward [25] and emotion [26] in relation to the self, where participants associate a geometric shape with a particular reward value or a particular emotion. In addition, the perceptual matching paradigm benefits from other advantages including its test-retest reliability [27] and simplicity [28], making it easy to administer in special populations such as children, older adults and clinical patients [24, 29, 30].

Self in neuroeconomics: self- and other- related values and choices.

A burgeoning literature has examined self- and other-related valuation and choices using tasks adapted from experimental economics in combination with neuroimaging, a field called neuroeconomics. One typical paradigm is the *trust game* (**Fig. 2A**) [31]. In this task, the 'investor' sends some money x to the 'trustee'. This amount is tripled to $3x$ (i.e. simulating profits) before it reaches the 'trustee'. The trustee then repays the investor an amount y ($y \in [0,3x]$) and this process is repeated. Unlike the perceptual matching paradigm, the trust game paradigm offers a clear temporally explicit dissociation between self vs. other processing due to its sequential nature (**Fig. 2B**).

Neuroimaging techniques have revealed the brain networks supporting the self- and other-related processing observed in the trust game. In healthy controls, the middle cingulate cortex is activated when participants make investment decisions as the 'investor' ('self phase'); and the anterior cingulate cortex (ACC) is active when participants play the 'trustee' role and observe the investor's decisions ('other phase') [32]. In addition, insula activity correlates with both the amount of investment when healthy controls play the 'investor' and the amount of offers when they play the 'trustee' [32]. In sharp contrast, patients with borderline personality disorder who played the 'investor' role displayed flat insula activity in response to offers received from the partner during the 'other' phase, yet comparable levels of insula activity when they are making an investment decision during the 'self' phase (**Fig. 2C;** [33]). Another study showed that participants with autism displayed diminished cingulate activity during the self-decision phase, but preserved normal levels of cingulate responses when processing others' decisions and when they play the 'trustee' in the trust game (**Fig. 2D**; [34]).

Other studies have used the tasks that involve making decisions for oneself and another person. Results from these studies suggest that midline structures such as the mPFC encode value signals [35] and choices [36] related to both the self and others. When one's own choices deviate from others' choices, the insula is commonly activated [37]. **Agency** also directly affects motivated behaviors even when making choices for oneself. For example, intrinsic motivation activates the anterior insula to a greater extent than extrinsically motivated behavior [38]. Additionally, choice agency (i.e. having control of choices) is considered valuable itself. People prefer having a choice, than not having a

choice, and show increased activity in the ventral striatum when given a choice [39, 40]. Agency also promotes persistence when confronted with setbacks; a process modulated by ventral striatum and vmPFC [41]. Taken together, these studies suggest that neuroeconomic paradigms offer a unique opportunity for computational modeling of self- and other-processing, as well as quantifying self-related deficits in psychopathology; value signals (e.g. processed in striatum) also contributes a source of self-related information.

========== Fig. 2 approximately here ==========

Self in embodied and interoceptive paradigms.

The importance of bodily signals in selfhood has also been extensively studied in recent studies. This line of research emphasizes that sensory (e.g. proprioceptive) and physiological (e.g. heart rate) signals coming from the body are crucial for the conscious awareness of feelings and 'self as object' [42-46]. One classical paradigm in this field is the **rubber hand illusion** in which one can acquire a false ownership of a rubber hand under multisensory (e.g. visuotactile) conflicts [47]. Based on this paradigm, Lenggenhager and colleagues developed a virtual reality paradigm to probe an **out-of-body experience** [48] in which participants can acquire the false ownership of a virtual fake body when their own bodies and the fake body are stroked simultaneously [48]. The behavioral index of these bodily illusions is usually the 'perceptual drift' – the distance from the subjects' own hand/body to the perceived hand/body. Neurally, sensorimotor areas and the insula have been commonly implicated (see[49] for a review). These studies

suggest that the sense of the self largely depends on the spatial boundaries of the body; by experimentally manipulating bodily information, the perception of the self can also be altered.

A closely related line of work focuses on interoception, or the physiological conditions of the body. A commonly used heartbeat detection paradigm, for example, asks the subjects to either judge the timing of [50] or simply pay attention to their heartbeats [51, 52]. Interoceptive awareness measured in these studies also correlated with emotional awareness [50-53]. Neurally, interoception is implemented in a set of brain regions including the thalamus, brainstem, hypothalamus, amygdala, insula, cingulate, and somatomotor areas [50-52]. Computationally, the awareness of one's own emotions and feelings can be realized through **interoceptive inference**, the approximate **Bayesian inference** about internal bodily states [44, 54]. Altered interoceptive inference may serve as a mechanism for disordered self processing and emotion in neuropsychiatric populations (e.g. [55]). Taken together, these results support the notion of an 'embodied self', consistent with the 'material me' idea proposed by William James [1, 56] as well as the notion of **embodied cognition** [44, 45].

Self in social neuroscience: appraisal, theory of mind, and empathy.

People often form beliefs about themselves through the lens of others and interactions with others. Thus, self processing has also been extensively examined in the growing field of **social neuroscience**. Many of these studies have examined the neural correlates

of self and other representation using cognitive appraisal tasks, such as self-referential tasks [57]. In these tasks, participants are typically asked to make judgments about themselves and another person (e.g. "Am I / Is he brave?"). Consistent with findings from other self paradigms, these tasks commonly engage the mPFC, with a ventral-to-dorsal gradient for self- to other-related processing [57]. These tasks have also provided important insights into altered self processing in psychiatric disorders such as post-traumatic stress disorder [58], anorexia nervosa [59], and borderline personality disorder [60].

Previous work has also investigated the self in the context of **theory of mind** (ToM), the ability to infer others' thoughts and beliefs [61]. ToM is closely related to perspective taking (e.g. measured by the 'Sally-Ann test' [62]). It has been suggested that both self and other perspectives activate the mPFC and superior temporal gyrus [63]. Thus, it is not surprising that impairments in one's own awareness (e.g. as seen in **alexithymia)** are associated with impaired mentalizing and perspective-taking, as well as reduced activation in the mPFC [64]. Similarly, in conditions involving social cognitive impairments (e.g. autism), there is usually disturbed self processing [65, 66].

Lastly, existing studies have demonstrated that the representation of self-emotions is crucial for **empathy**, the representation of others' emotional states. Wicker and colleagues, for example, were amongst the first to observe common neural activation in the insula associated with feeling disgust oneself and observing others' disgust [67]. This

finding has since inspired many studies to examine shared neural representations for self- and empathetic emotions including pain, positive or negative emotions [68, 69]. These studies demonstrate that an insula–ACC network shows the most consistent activation in both the perception of one's own pain and the perception of others' pain [43, 70]. As an important region for emotion encoding, the amygdala also responds to both self and other emotions, as revealed by recent meta-analyses [43, 70]. In addition, the mPFC is activated by the conscious evaluation of self and other emotions [69]. Importantly, the disruption in processing self emotions has been directly linked to deficits in perceiving others' emotions in people with autism [55, 71]. These data collectively support a critical role of self processing in normal social and affective functions. While the dependency between self and other representations remains to be further investigated, these results clearly indicate a strong link between self and other perception in both cognitive and affective domains, and also provide quantitative paradigms for measuring the self through the reflection of others.

It is noteworthy that in addition to these new paradigms, emerging analytic techniques applied to 'old' methods have also yielded fruitful results. For instance, Lutz and colleagues divided participants' first-person reports into 'phenomenological clusters', and found that different, yet stable neurophysiological (EEG) signatures emerged for different phenomenological clusters [72]. Furthermore, techniques such as the Explicitation Interview (a guided retrospective introspection) and Descriptive Experience Sampling (in which subjects carry a beeper around and document their inner thoughts whenever the

beeper beeps) also offer new insights into our inner experiences through structured interviews and examinations [73].

Neural model of the self as object

Based on the literature reviewed above, we propose a neural framework of the 'self as object', which considers the self as an emerging property of interactions between brain networks implementing the 'core self', cognitive control, and salience processing (**Fig. 3**) [28]. These interactions across the networks reflect the relations between the strengths of cognitive representation and emotional response related to the self.

========== Insert Fig. 3 ==========

This neural model is built on the integrative property of the self. At the behavioral level, self-reference serves a binding function to facilitate various stages of information processing [15, 17, 27]. Neurally, the integrative property of the self is associated with the function of the vmPFC extending to the ACC [74], and how these areas couple with other brain regions [75]. There is a gradient of self-related processing in the ventral-dorsal axis along the mPFC [57] - while activity in the vmPFC is consistently associated with internal self processing, the dorsal mPFC (dmPFC) is more engaged in other-related judgments [57] .

fMRI studies of self-processing have consistently demonstrated increased coupling between the vmPFC and brain regions important for cognitive control such as the pSTS and dlPFC [20]. The pSTS is thought to be critical for sensory integration and stimulus-driven social attention [76]. Researchers have reported increased functional coupling between vmPFC and pSTS during self processing [20], which contributes to a form of 'social saliency' in the presence of self-related stimuli. It has been also shown that the vmPFC and dlPFC have opposite response patterns in the presence of self-related stimuli: vmPFC shows increased activity, while dlPFC has decreased activity [20, 77]. In addition, dlPFC is more active when participants respond to stimuli associated with other people [20]. These contrasting effects support a processing network in which self-related neural circuits (e.g., vmPFC) interact with brain regions concerned with bottom-up (e.g. pSTS) and top-down cognitive control (e.g. dlPFC) (**Fig. 3**) [77]. When self-related information is irrelevant to the task, the activities of these regions may be set in opposition so that cognitive control regions must suppress self-related attention. Consistent with this, lesions affecting top-down attentional regions (e.g. dlPFC) can result in enhanced self-biases due to the release of attentional control [24, 78].

Interactions between the mPFC and the salience network have also been reported. Anatomically, there are rich connections between the insula, amygdala, striatum, and the mPFC [79, 80]. During rest, functional activity of the vmPFC and insula/amygdala/striatum, are anticorrelated – the former is considered as part of the 'default mode network' while the latter is considered part of the 'salience network' [81]. Lesion studies have shown that brain damage to both the vmPFC and the insula is

associated with reduced self-biases in perception and memory [24]. Disrupted insula-amygdala-mPFC connectivity has been often observed in psychiatric disorders [82]. It remains to be further investigated, using effective connectivity modeling (e.g. dynamic causal modeling [83]), how these regions causally interact during self processing.

This neural model of the self as object has broad potential implications for understanding a wide range of neuropsychological and neuropsychiatric disorders related to self-processing, that can be characterized by the extent of shifts away from a normative mean with the networks - the mPFC, dlPFC, pSTS, striatum, and sulcus (**Fig. 3**).

Concluding Remarks

We have proposed that by adopting the 'self as object' framework, it is possible to provide quantitative measures to characterize the self using emerging paradigms such as perceptual matching, the trust game, embodied self, and social neuroscience paradigms. This emerging objective measure of self biases that can be used as a proxy for self-representation is supported by a proposed neural framework which hypothesizes interactions between brain networks responding to the core self, cognitive control, and salience. There are many remaining issues to be addressed (see Outstanding Questions). For instance, what algorithm does the brain use to 'compute' the self? What are the causal interactions between the proposed neural regions? What neurotransmitters are involved in self processing? To the best of our knowledge, none of these questions has been formally addressed in the self literature. Importantly, it would be valuable to consider the self as a dimensional measure and apply the 'self as object' framework to

different clinical populations, as proposed by the Research Domain Criteria framework of the National Institute of Mental Health in the U.S. (**Box 1**). For example, bias in self-focused attention and affect is evident in several clinical populations such as depressed individuals. Thus, the paradigms and models reviewed here could potentially provide valuable insights into the mechanisms of mental illness.

Glossary

Alexithymia: A condition marked by impaired awareness of self emotions.

Bayesian inference: A statistical inference method in which Bayes' theorem is used to update the probability for a hypothesis as one accumulates more evidence or information.

Computational psychiatry: An emerging interdisciplinary field that seeks to characterize mental disorders in terms of aberrant computations at multiple scales.

Embodied cognition (embodiment): A theory that suggests that mental processes are shaped by aspects of the body.

Empathy: The understanding of other people's emotional states.

Empirical manipulation: A scientific method that involves the systematic control and change of experimental conditions in laboratory settings.

Event-related potential (ERP): Electrical activities of the brain due to a specific sensory, cognitive, or motor event, measured by eletroencephalography (EEG).

Interoception: The sense of the physiological states of the body. It is considered to be processed in a specific neural pathway that includes the thalamus, hypothalamus, insula, amygdala, and several other regions.

Interoceptive inference: the approximate **Bayesian inference** about internal bodily states.

Integrative self: A mechanistic function of self where self-reference enhances the binding of information and psychological processes through neural couplings between the

vmPFC and other brain regions. It can be used to interpret a wide range of optimal behaviors.

Interoceptive inference: The approximate Bayesian inference about internal bodily states.

Neuroeconomics: An interdisciplinary field that aims to understand the neural basis of economic decision-making, and how economic behavior can shape the brain.

Out-of-body experience: The sensation of separation from one's own body and seeing one's own body from a distance.

Rubber hand illusion: A visuotactile illusion in which the participant feels ownership of a fake rubber hand placed in front of her after watching the fake hand and her own hand being stroked at the same time.

Social neuroscience: An interdisciplinary field that seeks to investigate the neural basis of social processes.

Theory of Mind: The ability to infer and understand another person's thoughts, beliefs, and goals.

Reference

1. James, W. (1890) The principles of psychology, Macmillan.
2. Freud, S. and Riviere, J. (1947) The ego and the id, 4th.ed. edn., Hogarth Press ; Institute of Psycho-Analysis.
3. Dennett, D.C. (1992) The self as a center of narrative gravity. In Self and consciousness: Multiple perspectives, Hillsdale, NJ: Erlbaum.
4. Hood, B., The Self Illusion : Why There is No 'You' Inside Your Head, Constable, London, 2012, p. 1 online resource (160 pages).
5. Northoff, G. (2011) Self and brain: what is self-related processing? Trends Cogn Sci 15 (5), 186-7; author reply 187-8.
6. Rosenberg, M. (1965) Society and the adolescent self-image, Princeton university press Princeton, NJ.
7. Singelis, T.M. (1994) The measurement of independent and interdependent self-construals. Personality and social psychology bulletin 20 (5), 580-591.
8. Sui, J. et al. (2015) Super-Capacity Me! Super-Capacity and Violations of Race Independence for Self- but Not for Reward-Associated Stimuli. Journal of Experimental Psychology-Human Perception and Performance 41 (2), 441-452.
9. Kim, K. and Johnson, M. (2014) Extended self: spontaneous activation of medial prefrontal cortex by objects that are 'mine'. Social Cognitive and Affective Neuroscience 9 (7), 1006-1012.
10. Christoff, K. et al. (2011) Specifying the self for cognitive neuroscience. Trends Cogn Sci 15 (3), 104-12.
11. Montague, P.R. et al. (2012) Computational psychiatry. Trends in cognitive sciences 16 (1), 72-80.
12. Sui, J. and Humphreys, G. (2015) Super-size me: self biases increase to larger stimuli. Psychonomic Bulletin & Review 22 (2), 550-558.
13. Conway, M. and Pleydell-Pearce, C. (2000) The construction of autobiographical memories in the self-memory system. Psychological Review 107 (2), 261-288.
14. Ma, Y. and Han, S. (2010) Why we respond faster to the self than to others? An implicit positive association theory of self-advantage during implicit face recognition. J Exp Psychol Hum Percept Perform 36 (3), 619-33.
15. Macrae, C.N. et al. (2017) Self-relevance prioritizes access to visual awareness. J Exp Psychol Hum Percept Perform 43 (3), 438-443.
16. Sui, J. et al. (2012) Perceptual Effects of Social Salience: Evidence From Self-Prioritization Effects on Perceptual Matching. Journal of Experimental Psychology-Human Perception and Performance 38 (5), 1105-1117.
17. Sui, J. (2015) The integrative self: How self-reference integrates perception and memory. Trends in Cognitive Sciences.
18. Keyes, H. (2012) Categorical perception effects for facial identity in robustly represented familiar and self-faces: The role of configural and featural information. Quarterly Journal of Experimental Psychology 65 (4), 760-772.
19. Sui, J. et al. (2013) The Salient Self: The Left Intraparietal Sulcus Responds to Social as Well as Perceptual-Salience After Self-Association. Cerebral Cortex 25 (4), 1060-1068.

20. Sui, J. et al. (2013) Coupling social attention to the self forms a network for personal significance. Proceedings of the National Academy of Sciences of the United States of America 110 (19), 7607-7612.
21. Liu, M. et al. (2015) Dynamically orienting your own face facilitates the automatic attraction of attention. Cognitive Neuroscience.
22. Northoff, G. (2016) Is the self a higher-order or fundamental function of the brain? The "basis model of self-specificity" and its encoding by the brain's spontaneous activity. Cognitive neuroscience 7 (1-4), 203-222.
23. Araujo, H.F. et al. (2014) Involvement of cortical midline structures in the processing of autobiographical information. PeerJ 2, e481.
24. Sui, J., Enock, F., Ralph, J., and Humphreys, G.W. (2015) Dissociating hyper- and hypo-self biases to a core self-representation. Cortex 70, 202-212.
25. Sui, J. and Humphreys, G. (2015) The interaction between self-bias and reward: Evidence for common and distinct processes. Quarterly Journal of Experimental Psychology 68 (10), 1952-1964.
26. Stolte, M. et al. (2015) Dissociating Biases Towards the Self and Positive Emotion. Q J Exp Psychol (Hove), 1-34.
27. Wang, H. et al. (2015) Expanding and retracting from the self: Gains and costs in switching self-associations. Journal of Experimental Psychology: Human Perception and Performance.
28. Humphreys, G. and Sui, J. (2015) The salient self: Social saliency effects based on self-bias. Journal of Cognitive Psychology 27 (2), 129-140.
29. Sui, J. and Humphreys, G.W. (2017) Aging enhances cognitive biases to friends but not the self. Psychonomic Bulletin & Review, 1-10.
30. Sui, J. and Humphreys, G.W. (2017) The self survives extinction: Self-association biases attention in patients with visual extinction. Cortex.
31. Camerer, C. (2003) Behavioral game theory: Experiments in strategic interaction, Princeton University Press.
32. King-Casas, B. et al. (2005) Getting to know you: reputation and trust in a two-person economic exchange. Science 308 (5718), 78-83.
33. King-Casas, B. et al. (2008) The rupture and repair of cooperation in borderline personality disorder. Science 321 (5890), 806-10.
34. Chiu, P.H. et al. (2008) Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. Neuron 57 (3), 463-73.
35. Garvert, M.M. et al. (2015) Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. Neuron 85 (2), 418-28.
36. Nicolle, A. et al. (2012) An agent independent axis for executed and modeled choice in medial prefrontal cortex. Neuron 75 (6), 1114-21.
37. Wu, H. et al. (2016) Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. Neurosci Biobehav Rev 71, 101-111.
38. Lee, W. and Reeve, J. (2013) Self-determined, but not non-self-determined, motivation predicts activations in the anterior insular cortex: an fMRI study of personal agency. Soc Cogn Affect Neurosci 8 (5), 538-45.
39. Leotti, L.A. and Delgado, M.R. (2011) The inherent reward of choice. Psychol Sci 22 (10), 1310-8.

40. Leotti, L.A. and Delgado, M.R. (2014) The value of exercising control over monetary gains and losses. Psychol Sci 25 (2), 596-604.
41. Bhanji, J.P. and Delgado, M.R. (2014) Perceived control influences neural responses to setbacks and promotes persistence. Neuron 83 (6), 1369-75.
42. Craig, A.D. (2014) How Do You Feel?: An Interoceptive Moment with Your Neurobiological Self, Princeton University Press.
43. Gu, X. et al. (2013) Anterior insular cortex and emotional awareness. J Comp Neurol 521 (15), 3371-88.
44. Seth, A. (2013) Interoceptive inference, emotion, and the embodied self. Trends in Cognitive Sciences 17 (11), 565-573.
45. Niedenthal, P.M. (2007) Embodying emotion. Science 316 (5827), 1002-5.
46. Damasio, A. (2008) Descartes' error: Emotion, reason and the human brain, Random House.
47. Botvinick, M. and Cohen, J. (1998) Rubber hands' feel'touch that eyes see. Nature 391 (6669), 756.
48. Lenggenhager, B. et al. (2007) Video ergo sum: manipulating bodily self-consciousness. Science 317 (5841), 1096-9.
49. Blanke, O. (2012) Multisensory brain mechanisms of bodily self-consciousness. Nat Rev Neurosci 13 (8), 556-71.
50. Critchley, H.D. et al. (2004) Neural systems supporting interoceptive awareness. Nat Neurosci 7 (2), 189-95.
51. Pollatos, O. et al. (2007) Brain structures mediating cardiovascular arousal and interoceptive awareness. Brain Res 1141, 178-87.
52. Zaki, J. et al. (2012) Overlapping activity in anterior insula during interoception and emotional experience. Neuroimage 62 (1), 493-9.
53. Pollatos, O. et al. (2007) Heart rate response after emotional picture presentation is modulated by interoceptive awareness. Int J Psychophysiol 63 (1), 117-24.
54. Gu, X. and FitzGerald, T.H. (2014) Interoceptive inference: homeostasis and decision-making. Trends Cogn Sci 18 (6), 269-70.
55. Gu, X. et al. (2015) Autonomic and brain responses associated with empathy deficits in autism spectrum disorder. Hum Brain Mapp 36 (9), 3323-38.
56. James, W. (1884) What is an emotion? Mind os-IX (34), 188-205.
57. Denny, B.T. et al. (2012) A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. J Cogn Neurosci 24 (8), 1742-52.
58. Bluhm, R.L. et al. (2012) Neural correlates of self-reflection in post-traumatic stress disorder. Acta Psychiatr Scand 125 (3), 238-46.
59. McAdams, C.J. et al. (2016) Neural differences in self-perception during illness and after weight-recovery in anorexia nervosa. Soc Cogn Affect Neurosci 11 (11), 1823-1831.
60. Winter, D. et al. (2015) Negative evaluation bias for positive self-referential information in borderline personality disorder. PLoS One 10 (1), e0117083.
61. Shamay-Tsoory, S.G. et al. (2009) Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. Brain 132 (Pt 3), 617-27.

62. Wimmer, H. and Perner, J. (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13 (1), 103-28.
63. van Veluw, S.J. and Chance, S.A. (2014) Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. Brain imaging and behavior 8 (1), 24-38.
64. Moriguchi, Y. et al. (2006) Impaired self-awareness and theory of mind: an fMRI study of mentalizing in alexithymia. Neuroimage 32 (3), 1472-82.
65. Silani, G. et al. (2008) Levels of emotional awareness and autism: an fMRI study. Soc Neurosci 3 (2), 97-112.
66. Bird, G. et al. (2010) Empathic brain responses in insula are modulated by levels of alexithymia but not autism. Brain 133 (Pt 5), 1515-25.
67. Wicker, B. et al. (2003) Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. Neuron 40 (3), 655-64.
68. Zaki, J. et al. (2016) The Anatomy of Suffering: Understanding the Relationship between Nociceptive and Empathic Pain. Trends Cogn Sci 20 (4), 249-59.
69. Ochsner, K.N. et al. (2004) Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. J Cogn Neurosci 16 (10), 1746-72.
70. Lamm, C. et al. (2011) Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. Neuroimage 54 (3), 2492-502.
71. Gu, X. et al. (2017) Heightened Brain Response to Pain Anticipation in High-Functioning Adults with Autism Spectrum Disorder. European Journal of Neuroscience.
72. Lutz, A. et al. (2002) Guiding the study of brain dynamics by using first-person data: synchrony patterns correlate with ongoing conscious states during a simple visual task. Proc Natl Acad Sci U S A 99 (3), 1586-91.
73. Froese, T. et al. (2011) Validating and calibrating first-and second-person methods in the science of consciousness. Journal of Consciousness Studies 18 (2), 38.
74. Northoff, G. and Bermpohl, F. (2004) Cortical midline structures and the self. Trends Cogn Sci 8 (3), 102-7.
75. Sui, J. (2016) Self-Reference Acts as a Golden Thread in Binding. Trends in Cognitive Sciences 20 (7), 482-483.
76. Allison, T. et al. (2000) Social perception from visual cues: role of the STS region. Trends in Cognitive Sciences 4 (7), 267-278.
77. Humphreys, G.W. and Sui, J. (2015) Attentional control and the self: The Self-Attention Network (SAN). Cogn Neurosci, 1-13.
78. Sui, J. et al. (2013) Lesion-Symptom Mapping of Self-Prioritization in Explicit Face Categorization: Distinguishing Hypo- and Hyper-Self-Biases. Cerebral Cortex 25 (2), 374-383.
79. Flynn, F.G. et al. (1999) Anatomy of the insula - functional and clinical correlates. Aphasiology 13 (1), 55-78.
80. Kim, M.J. et al. (2011) The structural and functional connectivity of the amygdala: from normal emotion to pathological anxiety. Behav Brain Res 223 (2), 403-10.
81. Cauda, F. et al. (2011) Functional connectivity of the insula in the resting brain. Neuroimage 55 (1), 8-23.

82. McHugh, M.J. et al. (2014) Cortico-amygdala coupling as a marker of early relapse risk in cocaine-addicted individuals. Front Psychiatry 5, 16.
83. Friston, K.J. et al. (2003) Dynamic causal modelling. Neuroimage 19 (4), 1273-302.
84. Kozak, M.J. and Cuthbert, B.N. (2016) The NIMH Research Domain Criteria Initiative: Background, Issues, and Pragmatics. Psychophysiology 53 (3), 286-97.
85. Skodol, A. et al. (2011) Proposed Changes in Personality and Personality Disorder Assessment and Diagnosis for DSM-5 Part I: Description and Rationale. Personality Disorders-Theory Research and Treatment 2 (1), 4-22.
86. Northoff, G. (2007) Psychopathology and pathophysiology of the self in depression - neuropsychiatric hypothesis. J Affect Disord 104 (1-3), 1-14.
87. Groenewold, N.A. et al. (2013) Emotional valence modulates brain functional abnormalities in depression: evidence from a meta-analysis of fMRI studies. Neurosci Biobehav Rev 37 (2), 152-63.
88. Treadway, M.T. and Zald, D.H. (2011) Reconsidering anhedonia in depression: lessons from translational neuroscience. Neurosci Biobehav Rev 35 (3), 537-55.
89. Delaveau, P. et al. (2011) Brain effects of antidepressants in major depression: a meta-analysis of emotional processing studies. J Affect Disord 130 (1-2), 66-74.
90. Ma, Y. (2014) Neuropsychological mechanism underlying antidepressant effect: a systematic meta-analysis. Mol Psychiatry.
91. Beck, A.T. (1979) Cognitive therapy of depression, Guilford press.

Box 1. Self, Psychopathology, and the Research Domain Criteria (RDoC) era

RDoC is a recent initiative started by the NIMH, aiming to re-define the research framework for studying mental functions and dysfunctions [84]. The RDoC matrix integrates multiple levels of analyses (e.g. genetics, neural circuits) with functional domains (e.g. cognitive processes, social processes). Currently, there are five domains in RDoC: Negative Valence Systems, Positive Valence Systems, Cognitive Systems, Social Processes, and Arousal and Regulatory Systems. Importantly, perception and understanding of the self is listed as one of the four constructs under Social Processes (together with perception and understanding of others, social communication, and affiliation and attachment). Compared to many other construals in RDoC, reliable and objective instruments to measure the self in neuropsychiatric research are currently lacking, especially given the progress of studies on the self in psychology and cognitive neuroscience; and most measures of self processing in psychiatry have mainly depended on conventional qualitative self-report questionnaires [85].

To showcase the importance of quantifying the self with objective measures, we will focus on one example from the clinical realm: major depressive disorder (MDD). Quantitative paradigms have pointed at two self-related changes in depression. First, depressed patients usually show excessive self-focused attention [86]. It has been suggested that such a self-bias could divert one's attention from external stimuli, and is also highly related to rumination in depression [86]. Neurally, depressed patients show excessive mPFC activation but decreased dlPFC activation when making judgments about themselves in

self-referential tasks, providing a neural substrate for enhanced self-focused attention in these patients [86]. Second, studies using valence-based tasks typically suggest that for negative stimuli, MDD patients display decreased activation in dlPFC, yet hyper-activation in the amygdala and insula. For positive stimuli, however, patients show decreased insula and striatum activations but increased vmPFC activation, compared to controls [87]. These findings fit nicely with the anhedonia model of MDD [88], which considers MDD as a condition of reduced response to rewards (usually accompanied by reduced striatum activity). Taken together, these data confirm that MDD is marked by self-biased attention, and altered valence processing related to the self.

Although decreased or increased neural activations do not always directly correspond to reduced or enhanced behavioral sensitivity [20], we propose that tasks that can objectively measure the 'self as object' may be sensitive to clinical changes found in disorders such as MDD, and that classification using these objective measures of self prioritization may provide a way of moving diagnosis away from subjective evaluation. Studying self processing is also highly important for evaluating treatment outcomes. For example, antidepressant treatment can largely normalize disturbed neural responses to emotion (e.g. amygdala) in depressed patients [89, 90]. Cognitive behavioral therapy (CBT) is another mainstream treatment option, and one core component of CBT is the modification of maladaptive beliefs about the self [91]. Within the self attention framework [77], this modification (e.g. for MDD) can be conceptualized in terms of increased activation of the dorsal attention network compensating for decreased activation of the self-attentional network.

Figure Legends

Figure 1. Measuring the self using a perceptual matching task. A) Participants are instructed to associate three geometric shapes with three people, and then complete a shape-label matching task where shape-label pairs are presented as in their original assignment or not. **B)** The typical results are more accurate and faster responses to the self-associated shape compared to other-associated shapes, phenomena referred to as a self-prioritization effect. **C)** Self prioritization is supported by a specific neural circuit between the ventromedial prefrontal cortex (vmPFC) and left posterior superior temporal sulcus (pSTS). By contrast, matching stranger- compared to self-associated shapes recruits an executive control network. **D)** The ventral self network and the executive control network play opposite roles in self and other processing (Panels B-C-D are modified from [20]).

Figure 2. Measuring the self using a neuroeconomic paradigm: the trust game. A) In this version of the trust game, player one, the 'investor', sends a certain amount of money x out of an initial endowment of 20 monetary units (MU), to player two, the 'trustee'. This amount is tripled to 3x before it reaches the trustee. The trustee then repays the investor an amount y ($y \in [0,3x]$) and the process is repeated. **B)** In a typical session, the participant plays the role of either the investor (upper panel) or the trustee (lower panel). The 'self' vs. 'other' phases would thus be of opposite order for the two conditions. **C)** Subjects with borderline personality disorder (BPD) show reduced neural response in the 'other' phase but not in the 'self' phase when they played the role of investor (modified

from [33]). **D)** Subjects with autistic spectrum disorder (ASD) show diminished 'self' response but spared 'other' response when playing the role of trustee (modified from [34]). The bars in panels C-D are schematic, and do not quantitatively reflect actual data.

Figure 3. A neural model of the self as object. Self is implemented through a 'core self' network (e.g. mPFC), a cognitive control network (e.g. dlPFC, pSTS), and salience/affective network (e.g. insula, amygdala, striatum). The mPFC (extending into ACC) is considered to mediate internally-focused mental processing, which is usually inhibited in order to maintain attention on an on-going task [77]. There is a gradient of self-related processing in the ventral-dorsal axis along the mPFC [57] – the vmPFC is consistently associated with internal self processing whereas the dmPFC is more engaged in other-related judgments [57]. The dlPFC and pSTS are considered to process attention and cognitive control related to external stimuli. The insula, striatum, and amygdala are involved in the processing of salient emotional and reward stimuli. ACC: anterior cingulate cortex; pSTS: posterior superior temporal sulcus; dmPFC, vmPFC and dlPFC: dorso-medial, ventro-medial and dorso-lateral prefrontal cortex, respectively.

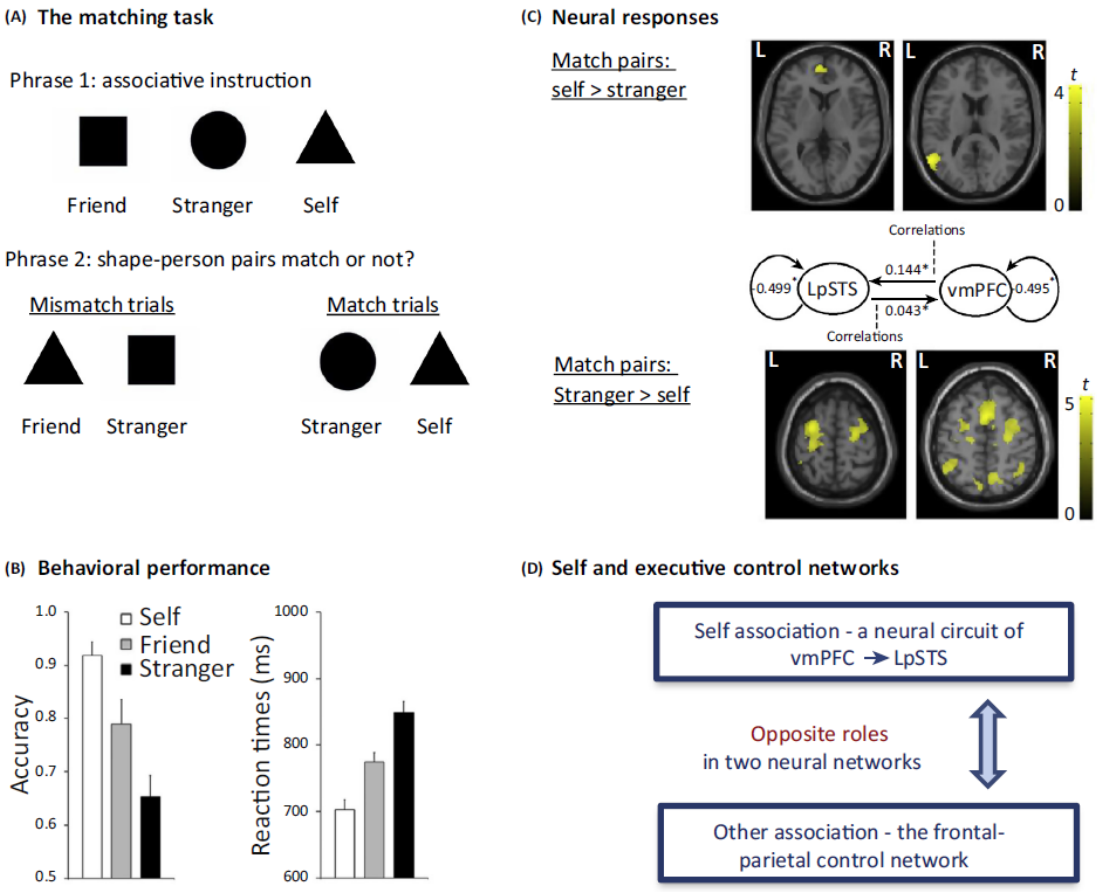
(A) The matching task
Phrase 1: associative instruction
Friend
Stranger
Self
Phrase 2: shape-person pairs match or not?
Mismatch trials
Friend
Stranger
Match trials
Stranger
Self
(B) Behavioral performance
Self
Friend
Stranger
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Reaction times (ms)
1000
900
800
700
600
(C) Neural responses
Match pairs:
self > stranger
L
R
t
4
0
Correlations
0.499*
LpSTS
0.144*
0.043*
vmPFC
0.495*
Correlations
Match pairs:
Stranger > self
5
(D) Self and executive control networks
Self association - a neural circuit of
vmPFC → LpSTS
Opposite roles
in two neural networks
Other association - the frontal-
parietal control network

Figure 1

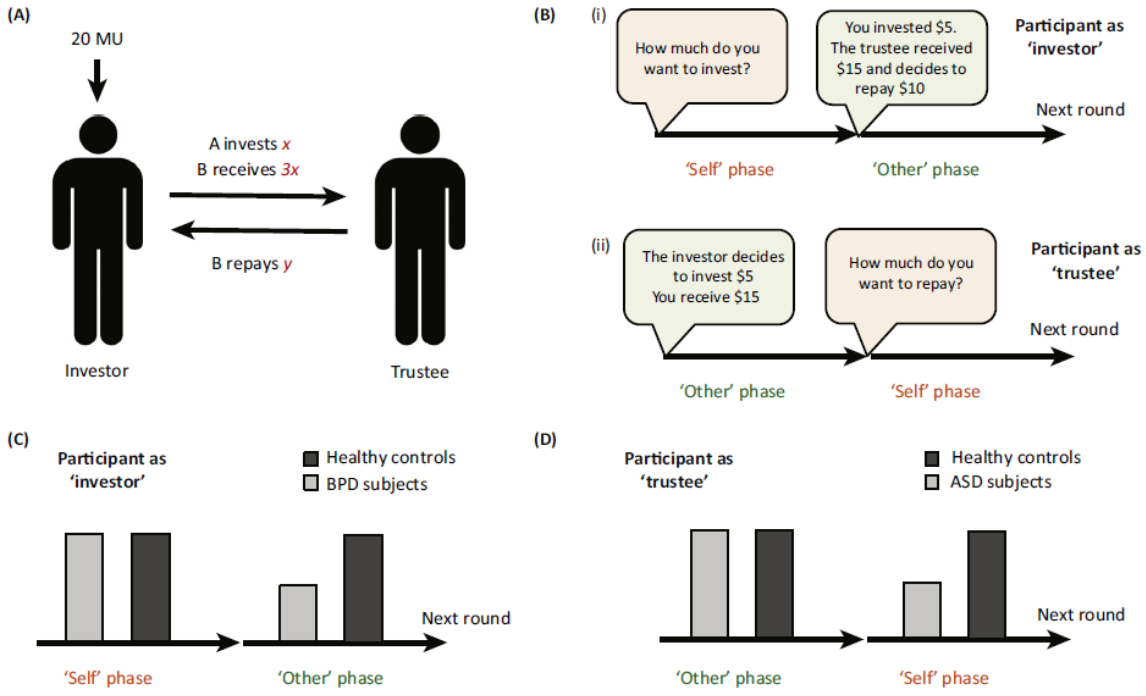
(A)
20 MU
A invests x
B receives 3x
B repays y
Investor
Trustee
(B)
(i)
How much do you want to invest?
You invested $5. The trustee received $15 and decides to repay $10
Participant as 'investor'
Next round
'Self' phase
'Other' phase
(ii)
The investor decides to invest $5 You receive $15
How much do you want to repay?
Participant as 'trustee'
Next round
'Other' phase
'Self' phase
(C)
Participant as 'investor'
Healthy controls
BPD subjects
Next round
'Self' phase
'Other' phase
(D)
Participant as 'trustee'
Healthy controls
ASD subjects
Next round
'Other' phase
'Self' phase

Figure 2

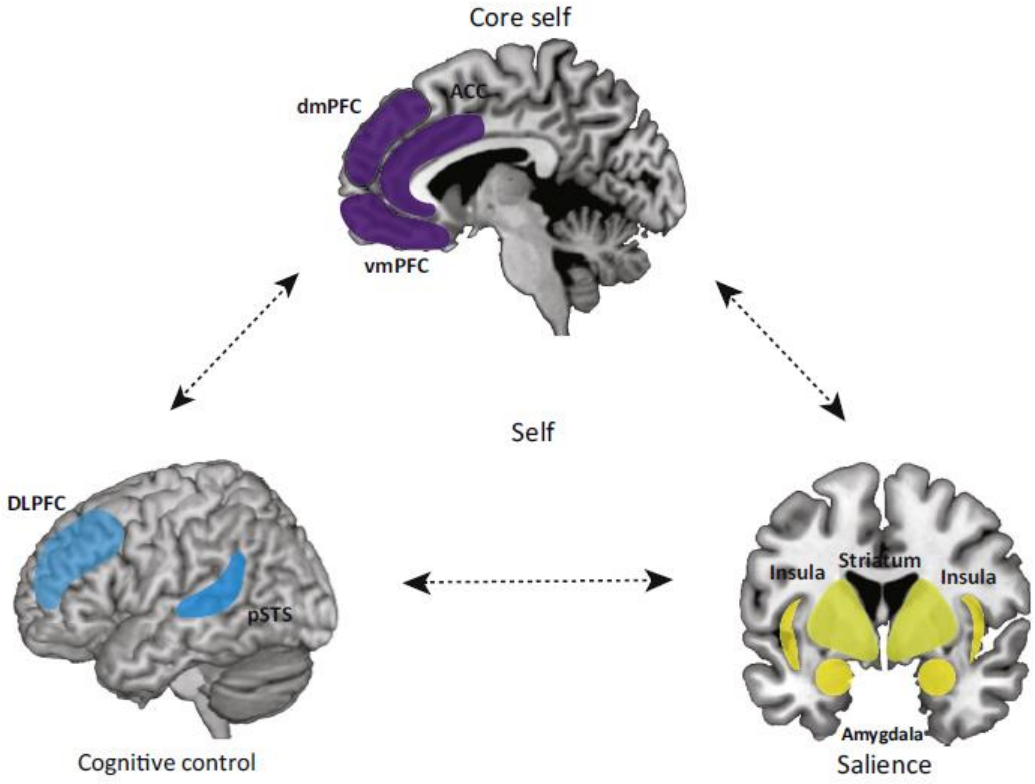
Core self
dmPFC
ACC
vmPFC
Self
DLPFC
pSTS
Cognitive control
Insula
Striatum
Insula
Amygdala
Salience

Figure 3